CONFIDENCE INTERVAL

→ First we should understand sampling and sampling error.

→ **Sample** – A sample is a selection of objects or observations taken from the population of interest.

→ **Example of sample** – A population might be all apples in a garden at a given time. We wish to measure weight of all the apple.

→ Inference is when we draw conclusion about the population from the sample. Suppose we took 3 batches/samples of apple. Sample mean Batch 1 = 149g, Batch 2 = 130g, Batch 3 = 153g

Difference in sample mean is called Sampling Error/Variation due to sampling. There will always be sampling error.

→ Normally whenever we give estimate of population based on sample, we should not say equals to but rather say lies between. For example, weight of the apple is in range between ___ with confidence interval ___.

→ What is the average weight of the apples in the garden?

Suppose sample mean = 149g.

We said that mean weight of the apples in the garden lies between 147 & 151g.

Factors affects the width of Confidence Intervals –

I) Variation → Variation within population of interest.
(In population)
- If all the values in the sample are same/almost similar the low variation
- Population with low variation leads to similar samples with low variation leads to narrow confidence interval.
- But a more varied population will lead to more varied sample
- Population with lots of variation leads to varied samples with high variation leads to wider confidence interval.

II) Sample size → Sample size also affects the width of a confidence interval
- If we take a small sample size, we don't have much information of base for our inference.

→ More samples will vary from each other, there will be more variation due to sampling or sampling error with small sample.

→ Large samples are more similar to each other and have more information, which leads to narrower confidence intervals.

Calculation →

According to Central limit theorem, $\text{Confidence Interval} = \bar{x} \pm t\frac{s}{\sqrt{n}}$

In traditional confidence intervals we wish to specify level of confidence. The more confidence we wish to be the larger our confidence interval will be. 90%, 95%, 99% confidence. 95%, we are 95% confident that the population mean lies within this interval.

In above example, avg mean of apple.
Sample size → 15, s → 4.758.
Standard error → $\frac{4.758}{\sqrt{15}} = 1.228$.
In t-table, n = 15, confidence interval = 95%
t value is 2.145 $= \bar{x} \pm 2.145(1.228)$
$= \bar{x} \pm 2.6$
$= 149 \pm 2.6 = (146.7, 151.9)$ with 95% confidence.
but 5% will be outside (146.7, 151.9)

s → Sample standard deviation
$\sqrt{n}$ → square root of Sample size
$s/\sqrt{n}$ → standard error
t → t distribution
Bigger t value, big confidence interval
$\bar{x}$ → sample mean

CENTRAL LIMIT THEOREM (CLT)

- In this, whatever the dataset its distribution does not matter. It can be uniform/ binomial / or any distribution.
- Take some sample (sufficient sample) and start finding mean of each sample.

Then use CLT,

No matter the distribution, extract mean of each sample set. Sample should be sufficient in number and bigger the sample, it will tend to follow normal distribution.

- Sample size should be greater than 25. Bigger the sample size, better the result.

Main Idea → Imagine a dataset with millions of values, and we can afford to sample small set and can be considered normal distribution.

$X \overset{\text{may/ may not}}{\not\sim} GD(\mu, \sigma^2)$

(Random Variable)

This may/may not belong to normal distribution/ Gaussian distribution.

Example - Consider 30 data points sample.

Sample size $\geq$ ~~30~~ 25.

Sample	Values.
S_1	$x_1, x_2, x_3 \ldots x_{30} = \bar{x}_1$
S_2	$\ldots\ldots x_{30} = \bar{x}_2$
S_3	$\ldots\ldots x_{30} = \bar{x}_3$
⋮	⋮
S_{100}	$= \bar{x}_{100}$

So, according to CLT if we plot $\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots \bar{x}_{100}$ it will follow a normal distribution.

And the mean of distribution will be $\bar{x}$ (mean of full set) or μ.

It will follow, $\boxed{\bar{x} = GD\left(\mu, \frac{\sigma^2}{n}\right)}$.